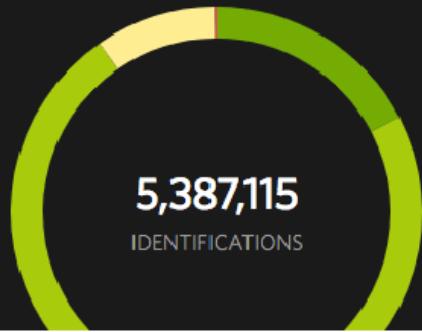


PREDICTING THE ACCURACY OF CITIZEN COLLECTED BIODIVERSITY DATA



Year In Review 2017

SHARE



The iNaturalist App

How It Works



1

Record your observations



2

Share with fellow naturalists



3

Discuss your findings

The iNaturalist Data

- Valuable biodiversity data collected by everyday people
- People share observations of plants and animals with putative identifications
- The community votes on the accuracy of the identifications

A screenshot of the iNaturalist mobile application interface. At the top, there is a header bar with icons for search, signal strength, time (1:07 PM), battery level (57%), and a gear icon. Below the header is a user profile card for "taniajogesh" featuring a circular profile picture of a person holding a lizard, the text "28 Observations", and a green gear icon. The main content area displays a list of recent observations, each with a small thumbnail image, the species name, a location, and a timestamp. The observations listed are: Townsend's Warbler (6740–6898 Skyline Blvd, Orinda, CA, 2y ago, 3 comments); Arboreal Salamander (Briones Regional Park, Lafayette, CA, 2y ago, 2 comments); Yellow-eyed Ensatina (Briones Regional Park, Lafayette, CA, 2y ago, 3 comments); California Slender Salamander (Briones Regional Park, Lafayette, CA, 2y ago, 2 comments); Broad-leaved Helleborine (Mount Tamalpais State Park, Mill Valley, CA, 2y ago, 1 comment); alpine false springparsley (32.952068, -108.209821, 2y ago); Butterfly Milkweed (Gila National Forest, Silver City, NM, 2y ago, 2 comments); Crevice Spiny Lizard (88061, Silver City, NM, US, 2y ago, 1 comment); and Cochise adder's-mouth orchid (2y ago). At the bottom of the screen are five navigation icons: "Explore" (globe), "Activity" (list), "Observe" (camera), "Me" (person silhouette with a green gear), and "More" (three dots).

Observation	Location	Timestamp	Comments
Townsend's Warbler	6740–6898 Skyline Blvd, Orinda, CA	2y	3
Arboreal Salamander	Briones Regional Park, Lafayette, CA	2y	2
Yellow-eyed Ensatina	Briones Regional Park, Lafayette, CA	2y	3
California Slender Salamander	Briones Regional Park, Lafayette, CA	2y	2
Broad-leaved Helleborine	Mount Tamalpais State Park, Mill Valley, CA	2y	1
alpine false springparsley	32.952068, -108.209821	2y	
Butterfly Milkweed	Gila National Forest, Silver City, NM	2y	2
Crevice Spiny Lizard	88061, Silver City, NM, US	2y	1
Cochise adder's-mouth orchid		2y	

The Problem

- Correctly identified data provides a treasure trove of information but species are often difficult to identify
- Can we use machine learning to tag observations that are likely to be correct?

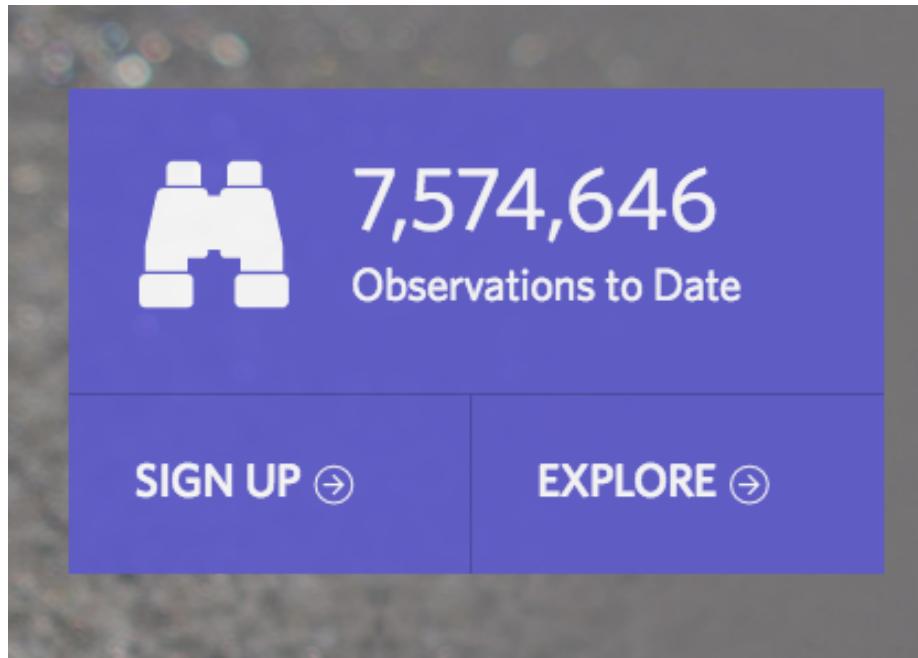
Activity

The screenshot shows a timeline of activity from the iNaturalist platform. It displays three separate identification suggestions for the same organism:

- taniajogesh suggested an ID** (3 years ago): Western Tiger Salamander (*Ambystoma mavortium*). Includes a "Compare" button.
- wild-about-texas suggested an ID** (1 year ago): Western Tiger Salamander (*Ambystoma mavortium*). Includes a "Compare" button.
- calebcam suggested an ID** (2 months ago): Arizona Tiger Salamander (*Ambystoma mavortium* ssp. *nebulosum*). Includes a "Compare" button and an "Agree" button.

At the bottom of the feed, there are buttons for "Comment" and "Suggest an Identification".

The Data



- Queried 10,000 observations using an API
- Limited to data collected between 2016-2017 in North America

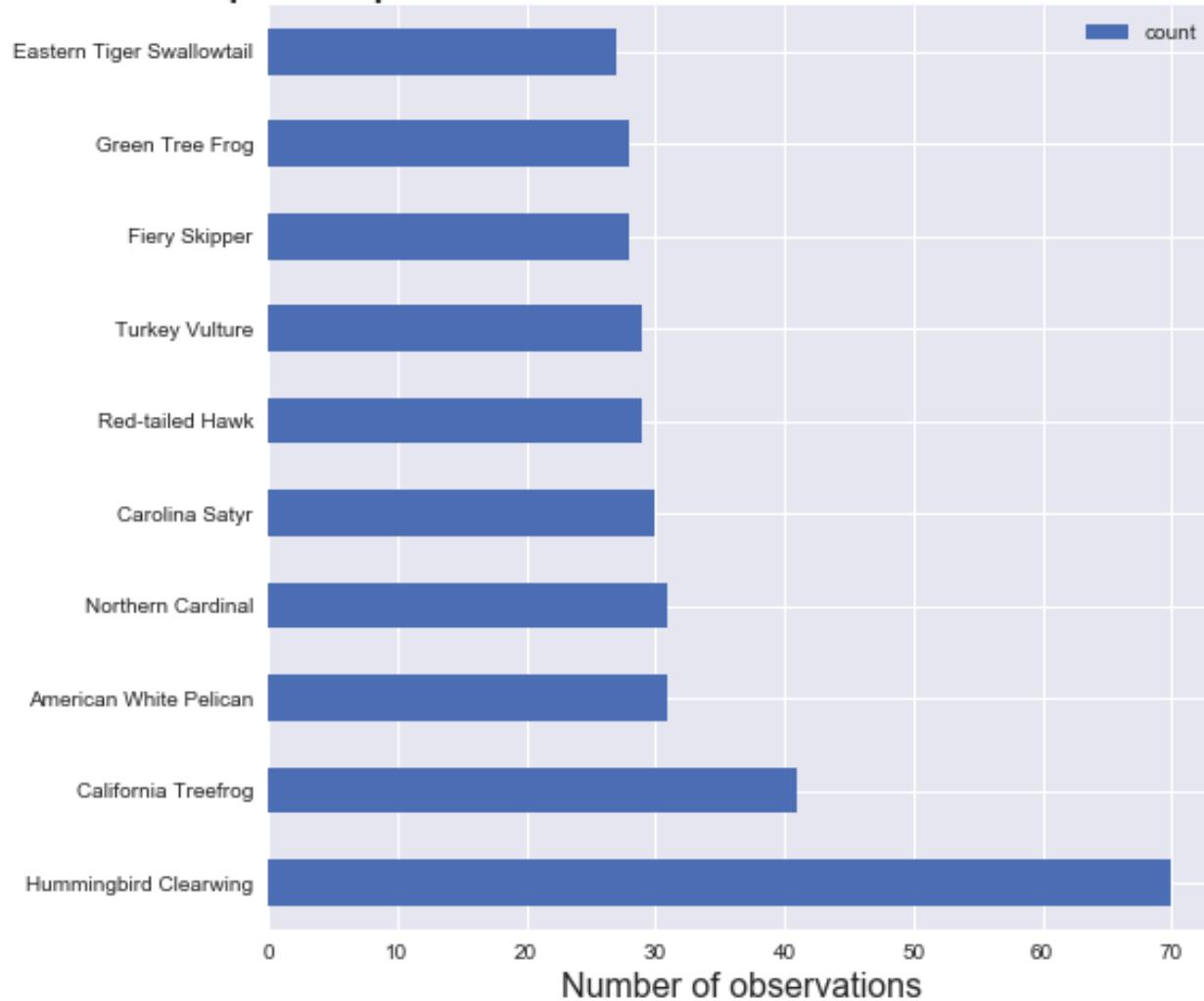
The data

- 450.9 MB of data
- Each observation as individual json file
- Spark Dataframes to read, parse and explore the data

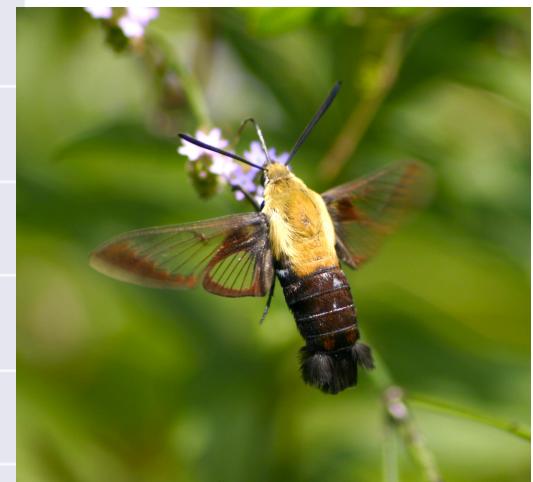


Top observations

Top 10 species observed on iNaturalist in the US



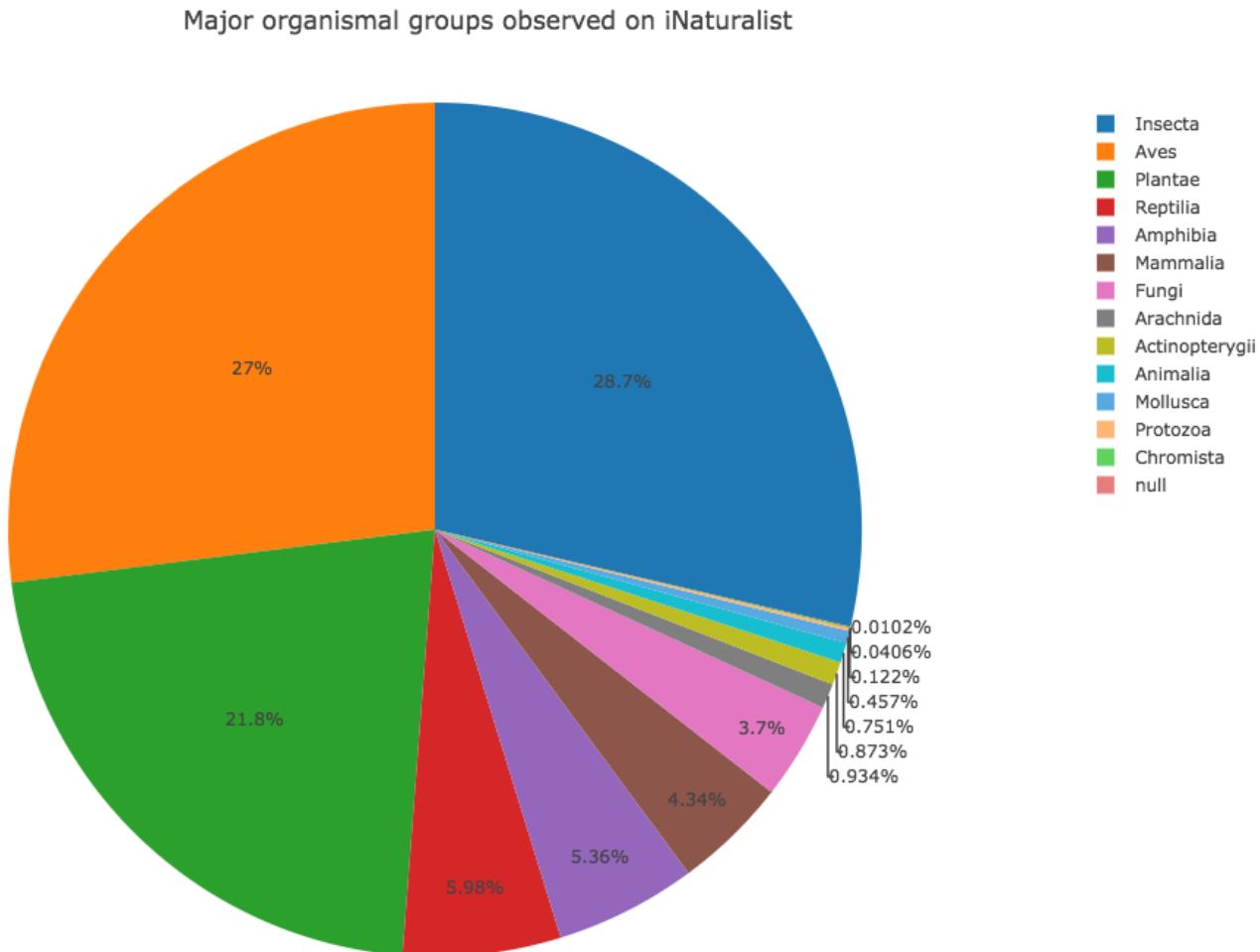
The data suggests that observations on iNaturalist are likely to be biased in favor of species that are more charismatic. Indeed, 80% of the top 10 list are birds and butterflies!



Hummingbird Clearwing

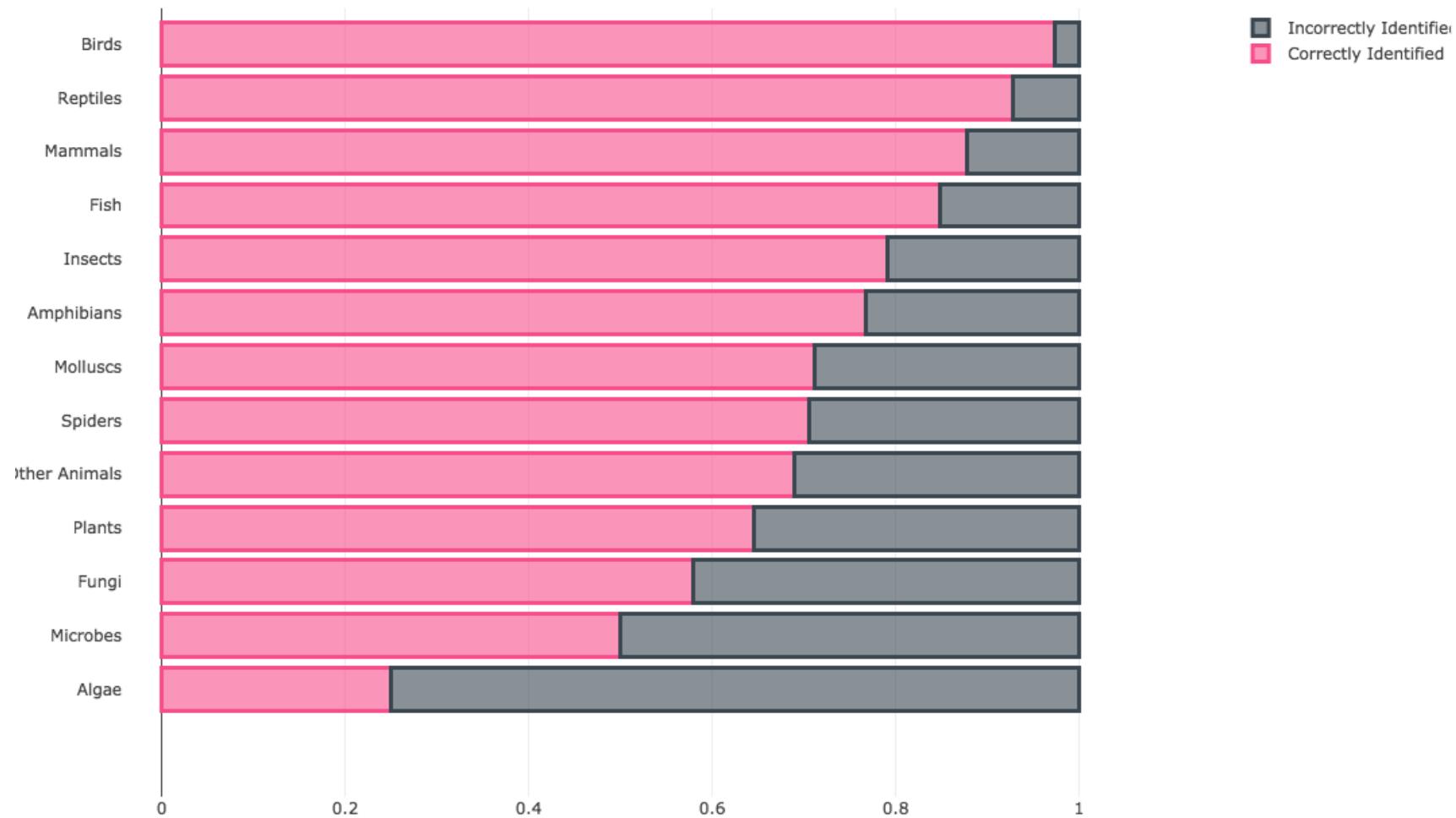
Major groups observed

Insects, Birds
and Plants
comprise the
bulk of
observations on
iNaturalist



Features correlated with identification accuracy

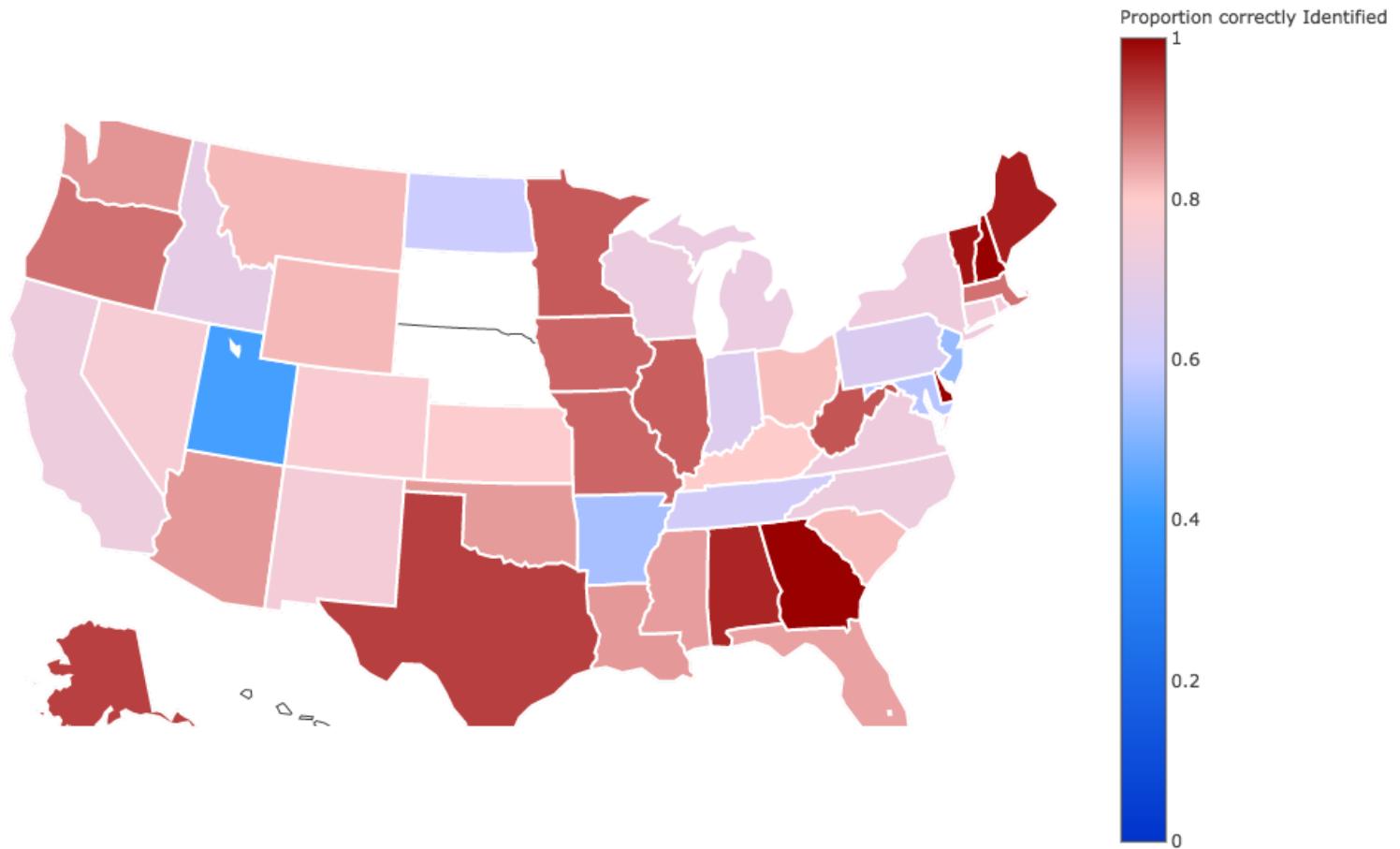
Taxonomic group



Features correlated with identification accuracy

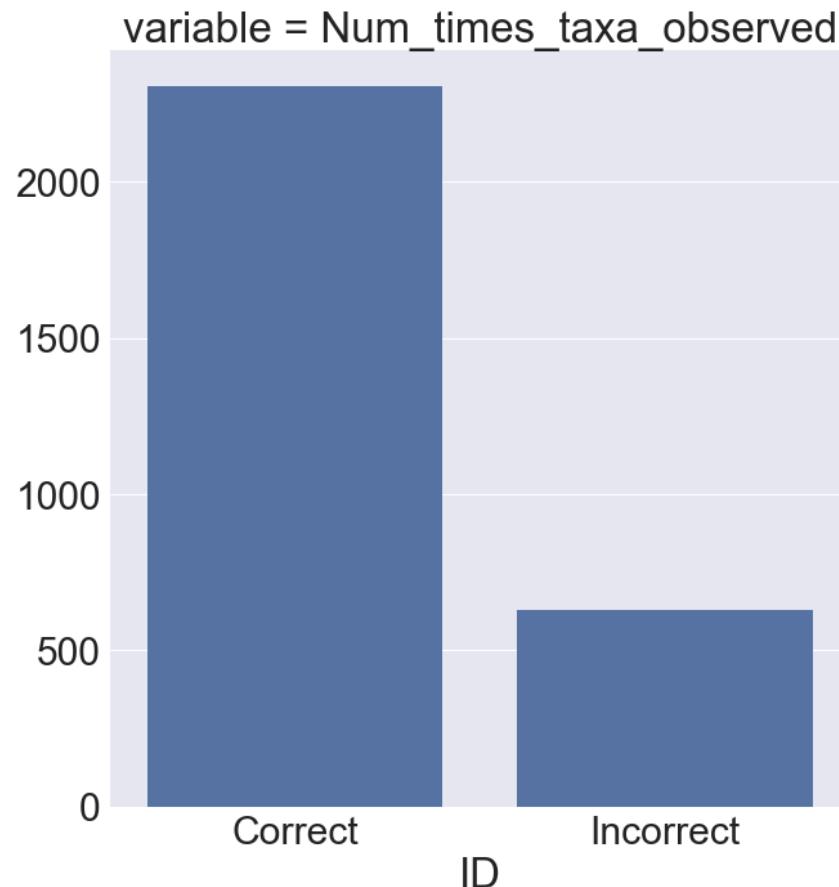
Location

iNaturalist accurate IDs by state



Features that predict identification accuracy

Number of times the species has been observed in the past is a strong predictor of accuracy



Feature engineering

How to treat categorical variables?

- 1) encode as numeric labels (not meaningful for linear methods)
- 2) one hot encoding to create a matrix with each category having a 0 or 1
- 3) scale features prior to ML

Machine Learning Models

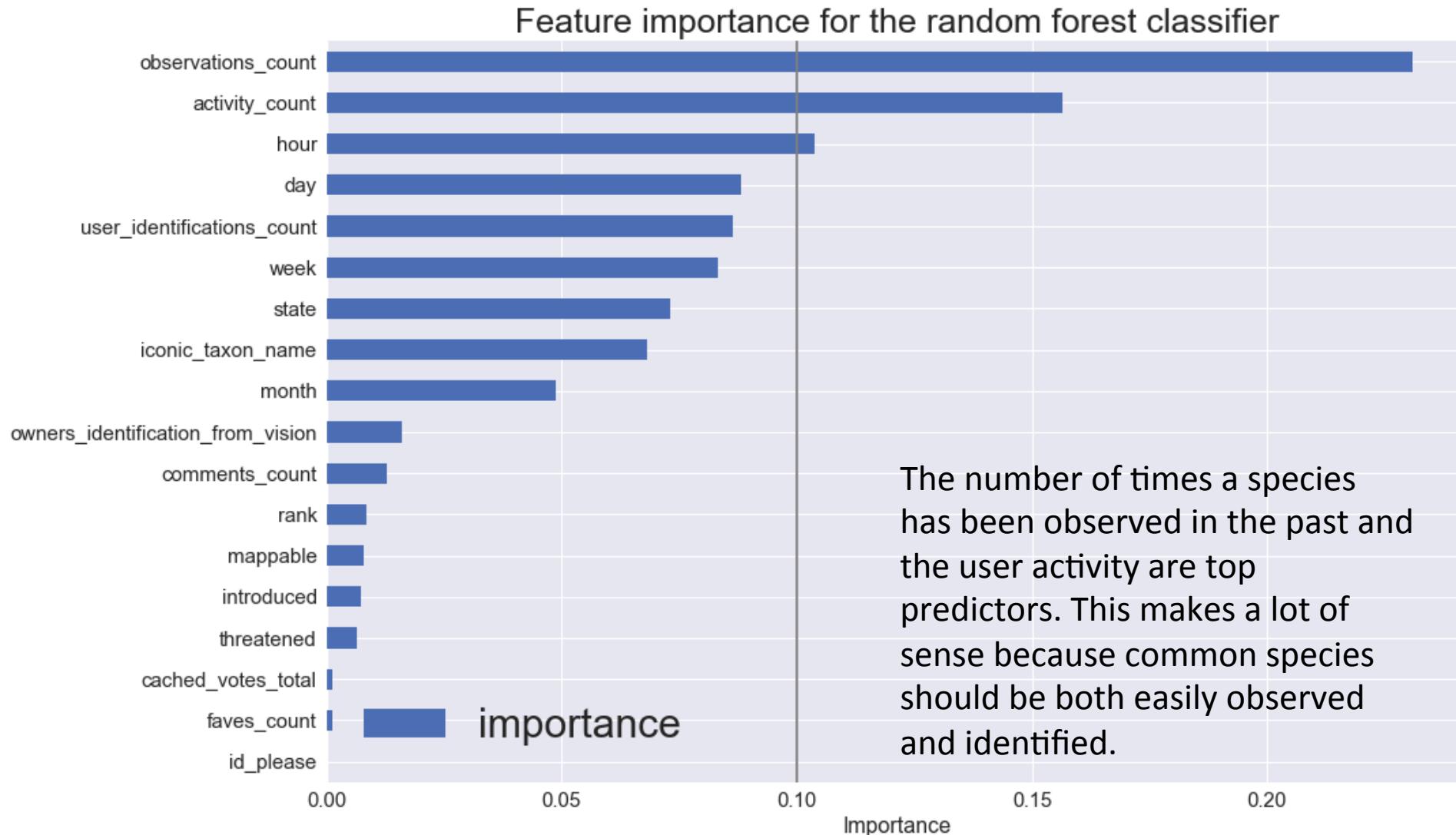
Model	optimized hyper-parameters	Categorical Features	Train accuracy	Test accuracy	speed
SVM	C=1, gamma =1	labeled	81.2%	82.1%	2.03s
		one hot encoded	87.4%	83.7%	21.86s
Random forest	default	labeled	99.7%	87.6%,	1.13s
		one hot encoded	99.7%	88.2%	1.99s
Gradient Boosting	max depth = 9, learning rate = 0.01, subsample = 0.6, min_leaf=1	labeled	99.7%	87.1%	33.62s
		one hot encoded	98.7%	87.4%	136.39s

Best classifier

Model	optimized hyper-parameters	Categorical Features	Train accuracy	Test accuracy	speed
Random forest		one hot encoded	99.7%	88.2% AUC = 0.89	1.99s
Gradient Boosting		one hot encoded	98.7%	87.4% AUC = 0.87	136.39s

The performance of the Random forest and Gradient Boosting classifier is pretty close. However, the Random Forest wins for being 100x faster with a better AUC score.

Top features that predict accuracy



Conclusions

- This classifier can be used by iNaturalist to automatically assign an accuracy rating to new observations.
- Observations with high accuracy can contribute to robust biodiversity data whereas, those predicted to be inaccurate can be tagged for expert identification